Exam : **CCA-500**

Title : Cloudera Certified Administrator for Apache Hadoop (CCAH)

Version : V9.02

1.Your cluster's mapred-start.xml includes the following parameters
<name>mapreduce.map.memory.mb</name>
<value>4096</value>
<name>mapreduce.reduce.memory.mb</name>
<value>8192</value>
And any cluster's yarn-site.xml includes the following parameters
<name>yarn.nodemanager.vmen-pmen-ration</name>
<value>2.1</value>
What is the maximum amount of virtual memory allocated for each map task before YARN will kill its Container?
A. 4 GB
B. 17.2 GB
C. 8.9GB
D. 8.2 GB
E. 24.6 GB

2.Assuming you're not running HDFS Federation, what is the maximum number of NameNode daemons you should run on your cluster in order to avoid a "split-brain" scenario with your NameNode when running HDFS High Availability (HA) using Quorum-based storage?
A. Two active NameNodes and two Standby NameNodes
B. One active NameNode and one Standby NameNode
C. Two active NameNodes and on Standby NameNode
D. Unlimited. HDFS High Availability (HA) is designed to overcome limitations on the number of NameNodes you can deploy

3.Table schemas in Hive are:
A. Stored as metadata on the NameNode
B. Stored along with the data in HDFS
C. Stored in the Metadata
D. Stored in ZooKeeper

4.For each YARN job, the Hadoop framework generates task log file. Where are Hadoop task log files stored?
A. Cached by the NodeManager managing the job containers, then written to a log directory on the NameNode
B. Cached in the YARN container running the task, then copied into HDFS on job completion
C. In HDFS, in the directory of the user who generates the job
D. On the local disk of the slave mode running the task

5.You have a cluster running with the fair Scheduler enabled. There are currently no jobs running on the

cluster, and you submit a job A, so that only job A is running on the cluster. A while later, you submit Job B. now Job A and Job B are running on the cluster at the same time. How will the Fair Scheduler handle these two jobs?
A. When Job B gets submitted, it will get assigned tasks, while job A continues to run with fewer tasks.
B. When Job B gets submitted, Job A has to finish first, before job B can gets scheduled.
C. When Job A gets submitted, it doesn't consumes all the task slots.
D. When Job A gets submitted, it consumes all the task slots.

6.Each node in your Hadoop cluster, running YARN, has 64GB memory and 24 cores. Your yarn.site.xml has the following configuration:
<property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value>32768</value>
</property>
<property>
<name>yarn.nodemanager.resource.cpu-vcores</name>
<value>12</value>
</property>
You want YARN to launch no more than 16 containers per node. What should you do?
A. Modify yarn-site.xml with the following property:
<name>yarn.scheduler.minimum-allocation-mb</name>
<value>2048</value>
B. Modify yarn-sites.xml with the following property:
<name>yarn.scheduler.minimum-allocation-mb</name>
<value>4096</value>
C. Modify yarn-site.xml with the following property:
<name>yarn.nodemanager.resource.cpu-vccores</name>
D. No action is needed: YARN's dynamic resource allocation automatically optimizes the node memory and cores

7.You want to node to only swap Hadoop daemon data from RAM to disk when absolutely necessary. What should you do?
A. Delete the /dev/vmswap file on the node
B. Delete the /etc/swap file on the node
C. Set the ram.swap parameter to 0 in core-site.xml
D. Set vm.swapfile file on the node
E. Delete the /swapfile file on the node

8.You are configuring your cluster to run HDFS and MapReducer v2 (MRv2) on YARN. Which two daemons needs to be installed on your cluster's master nodes?
A. HMaster

B. ResourceManager
C. TaskManager
D. JobTracker
E. NameNode
F. DataNode

9.You observed that the number of spilled records from Map tasks far exceeds the number of map output records. Your child heap size is 1GB and your io.sort.mb value is set to 1000MB. How would you tune your io.sort.mb value to achieve maximum memory to disk I/O ratio?
A. For a 1GB child heap size an io.sort.mb of 128 MB will always maximize memory to disk I/O
B. Increase the io.sort.mb to 1GB
C. Decrease the io.sort.mb value to 0 D. Tune the io.sort.mb value until you observe that the number of spilled records equals (or is as close to equals) the number of map output records.

10.You are running a Hadoop cluster with a NameNode on host mynamenode, a secondary NameNode on host mysecondarynamenode and several DataNodes.
Which best describes how you determine when the last checkpoint happened?
A. Execute hdfs namenode –report on the command line and look at the Last Checkpoint information
B. Execute hdfs dfsadmin –saveNamespace on the command line which returns to you the last checkpoint value in fstime file
C. Connect to the web UI of the Secondary NameNode (http://mysecondary:50090/) and look at the "Last Checkpoint" information
D. Connect to the web UI of the NameNode (http://mynamenode:50070) and look at the "Last Checkpoint" information

11. What does CDH packaging do on install to facilitate Kerberos security setup?
A. Automatically configures permissions for log files at&MAPRED_LOG_DIR/userlogs
B. Creates users for hdfs and mapreduce to facilitate role assignment
C. Creates directories for temp, hdfs, and mapreduce with the correct permissions
D. Creates a set of pre-configured Kerberos keytab files and their permissions
E. Creates and configures your kdc with default cluster values

12.You want to understand more about how users browse your public website. For example, you want to know which pages they visit prior to placing an order. You have a server farm of 200 web servers hosting your website. Which is the most efficient process to gather these web server across logs into your Hadoop cluster analysis?
A. Sample the web server logs web servers and copy them into HDFS using curl
B. Ingest the server web logs into HDFS using Flume
C. Channel these clickstreams into Hadoop using Hadoop Streaming

D. Import all user clicks from your OLTP databases into Hadoop using Sqoop
E. Write a MapReeeduce job with the web servers for mappers and the Hadoop cluster nodes for reducers

13.Which three basic configuration parameters must you set to migrate your cluster from MapReduce 1 (MRv1) to MapReduce V2 (MRv2)?
A. Configure the NodeManager to enable MapReduce services on YARN by setting the following property in yarn-site.xml:
<name>yarn.nodemanager.hostname</name>
<value>your_nodeManager_shuffle</value>
B. Configure the NodeManager hostname and enable node services on YARN by setting the following property in yarn-site.xml:
<name>yarn.nodemanager.hostname</name>
<value>your_nodeManager_hostname</value>
C. Configure a default scheduler to run on YARN by setting the following property in mapredsite.xml:
<name>mapreduce.jobtracker.taskScheduler</name>
<Value>org.apache.hadoop.mapred.JobQueueTaskScheduler</value>
D. Configure the number of map tasks per jon YARN by setting the following property in mapred:
<name>mapreduce.job.maps</name>
<value>2</value>
E. Configure the ResourceManager hostname and enable node services on YARN by setting the following property in yarn-site.xml:
<name>yarn.resourcemanager.hostname</name>
<value>your_resourceManager_hostname</value>
F. Configure MapReduce as a Framework running on YARN by setting the following property in mapred-site.xml:
<name>mapreduce.framework.name</name>
<value>yarn</value>

14.You need to analyze 60,000,000 images stored in JPEG format, each of which is approximately 25 KB. Because you Hadoop cluster isn't optimized for storing and processing many small files, you decide to do the following actions:
Group the individual images into a set of larger files
Use the set of larger files as input for a MapReduce job that processes them directly with python using Hadoop streaming.
Which data serialization system gives the flexibility to do this?
A. CSV
B. XML
C. HTML
D. Avro
E. SequenceFiles
F. JSON

15.Identify two features/issues that YARN is designated to address:
A. Standardize on a single MapReduce API
B. Single point of failure in the NameNode
C. Reduce complexity of the MapReduce APIs
D. Resource pressure on the JobTracker
E. Ability to run framework other than MapReduce, such as MPI
F. HDFS latency

16.Which YARN daemon or service monitors a Controller's per-application resource using (e.g., memory CPU)?
A. ApplicationMaster
B. NodeManager
C. ApplicationManagerService
D. ResourceManager

17.Which is the default scheduler in YARN?
A. YARN doesn't configure a default scheduler, you must first assign an appropriate scheduler class in yarn-site.xml
B. Capacity Scheduler
C. Fair Scheduler
D. FIFO Scheduler

18.Which YARN process run as "container 0" of a submitted job and is responsible for resource qrequests?
A. ApplicationManager
B. JobTracker
C. ApplicationMaster
D. JobHistoryServer
E. ResoureManager
F. NodeManager

19.Which scheduler would you deploy to ensure that your cluster allows short jobs to finish within a reasonable time without starting long-running jobs?
A. Complexity Fair Scheduler (CFS)
B. Capacity Scheduler
C. Fair Scheduler

D. FIFO Scheduler

20.Your cluster is configured with HDFS and MapReduce version 2 (MRv2) on YARN. What is the result when you execute: hadoop jar SampleJar MyClass on a client machine?
A. SampleJar.Jar is sent to the ApplicationMaster which allocates a container for SampleJar.Jar
B. Sample.jar is placed in a temporary directory in HDFS
C. SampleJar.jar is sent directly to the ResourceManager
D. SampleJar.jar is serialized into an XML file which is submitted to the ApplicatoionMaster

21.You are working on a project where you need to chain together MapReduce, Pig jobs. You also need the ability to use forks, decision points, and path joins. Which ecosystem project should you use to perform these actions?
A. Oozie
B. ZooKeeper
C. HBase
D. Sqoop
E. HUE

22.Which process instantiates user code, and executes map and reduce tasks on a cluster running MapReduce v2 (MRv2) on YARN?
A. NodeManager
B. ApplicationMaster
C. TaskTracker
D. JobTracker
E. NameNode
F. DataNode
G. ResourceManager

23.Cluster Summary:
45 files and directories, 12 blocks = 57 total. Heap size is 15.31 MB/193.38MB(7%)

Configured capacity	:	17.33GB
DFS Used	:	144KB
Non DFS Used	:	5.49GB
DFS Remaining	:	11.84GB
DFS Used %	:	0%
DFS Remaining %	:	68.32GB
Live Nodes	:	6
Dead Nodes	:	1
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	6

Refer to the above screenshot.
You configure a Hadoop cluster with seven DataNodes and on of your monitoring UIs displays the details shown in the exhibit.
What does the this tell you?
A. The DataNode JVM on one host is not active
B. Because your under-replicated blocks count matches the Live Nodes, one node is dead, and your DFS Used % equals 0%, you can't be certain that your cluster has all the data you've written it.
C. Your cluster has lost all HDFS data which had bocks stored on the dead DatNode
D. The HDFS cluster is in safe mode

24.Which two features does Kerberos security add to a Hadoop cluster?
A. User authentication on all remote procedure calls (RPCs)
B. Encryption for data during transfer between the Mappers and Reducers
C. Encryption for data on disk ("at rest")
D. Authentication for user access to the cluster against a central server
E. Root access to the cluster for users hdfs and mapred but non-root access for clients

25.Assuming a cluster running HDFS, MapReduce version 2 (MRv2) on YARN with all settings at their default, what do you need to do when adding a new slave node to cluster?
A. Nothing, other than ensuring that the DNS (or/etc/hosts files on all machines) contains any entry for the new node.
B. Restart the NameNode and ResourceManager daemons and resubmit any running jobs.
C. Add a new entry to /etc/nodes on the NameNode host.
D. Restart the NameNode of dfs.number.of.nodes in hdfs-site.xml

26.Which YARN daemon or service negotiations map and reduce Containers from the Scheduler, tracking

their status and monitoring progress?
A. NodeManager
B. ApplicationMaster
C. ApplicationManager
D. ResourceManager

27.During the execution of a MapReduce v2 (MRv2) job on YARN, where does the Mapper place the intermediate data of each Map Task?
A. The Mapper stores the intermediate data on the node running the Job's ApplicationMaster so that it is available to YARN ShuffleService before the data is presented to the Reducer
B. The Mapper stores the intermediate data in HDFS on the node where the Map tasks ran in the HDFS /usercache/&(user)/apache/application_&(appid) directory for the user who ran the job
C. The Mapper transfers the intermediate data immediately to the reducers as it is generated by the Map Task
D. YARN holds the intermediate data in the NodeManager's memory (a container) until it is transferred to the Reducer
E. The Mapper stores the intermediate data on the underlying filesystem of the local disk in the directories yarn.nodemanager.locak-DIFS

28.You suspect that your NameNode is incorrectly configured, and is swapping memory to disk. Which Linux commands help you to identify whether swapping is occurring?
A. free
B. df
C. memcat
D. top
E. jps
F. vmstat
G. swapinfo

29.On a cluster running CDH 5.0 or above, you use the hadoop fs –put command to write a 300MB file into a previously empty directory using an HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another use see when they look in directory?
A. The directory will appear to be empty until the entire file write is completed on the cluster
B. They will see the file with a ._COPYING_ extension on its name. If they view the file, they will see contents of the file up to the last completed block (as each 64MB block is written, that block becomes available)
C. They will see the file with a ._COPYING_ extension on its name. If they attempt to view the file, they will get a ConcurrentFileAccessException until the entire file write is completed on the cluster D. They will

see the file with its original name. If they attempt to view the file, they will get a ConcurrentFileAccessException until the entire file write is completed on the cluster

30.Which command does Hadoop offer to discover missing or corrupt HDFS data?
A. Hdfs fs –du
B. Hdfs fsck
C. Dskchk
D. The map-only checksum
E. Hadoop does not provide any tools to discover missing or corrupt data; there is not need because three replicas are kept for each data block

31.You are planning a Hadoop cluster and considering implementing 10 Gigabit Ethernet as the network fabric. Which workloads benefit the most from faster network fabric?
A. When your workload generates a large amount of output data, significantly larger than the amount of intermediate data
B. When your workload consumes a large amount of input data, relative to the entire capacity if HDFS
C. When your workload consists of processor-intensive tasks
D. When your workload generates a large amount of intermediate data, on the order of the input data itself

32.Your cluster is running MapReduce version 2 (MRv2) on YARN. Your ResourceManager is configured to use the FairScheduler. Now you want to configure your scheduler such that a new user on the cluster can submit jobs into their own queue application submission. Which configuration should you set?
A. You can specify new queue name when user submits a job and new queue can be created dynamically if the property yarn.scheduler.fair.allow-undecleared-pools = true
B. Yarn.scheduler.fair.user.fair-as-default-queue = false and yarn.scheduler.fair.allowundecleared-pools = true
C. You can specify new queue name when user submits a job and new queue can be created dynamically if yarn .schedule.fair.user-as-default-queue = false
D. You can specify new queue name per application in allocations.xml file and have new jobs automatically assigned to the application queue

33.A slave node in your cluster has 4 TB hard drives installed (4 x 2TB). The DataNode is configured to store HDFS blocks on all disks. You set the value of the dfs.datanode.du.reserved parameter to 100 GB. How does this alter HDFS block storage?
A. 25GB on each hard drive may not be used to store HDFS blocks
B. 100GB on each hard drive may not be used to store HDFS blocks
C. All hard drives may be used to store HDFS blocks as long as at least 100 GB in total is available on the node
D. A maximum if 100 GB on each hard drive may be used to store HDFS blocks

34.What two processes must you do if you are running a Hadoop cluster with a single NameNode and six DataNodes, and you want to change a configuration parameter so that it affects all six DataNodes.
A. You must modify the configuration files on the NameNode only. DataNodes read their configuration from the master nodes
B. You must modify the configuration files on each of the six SataNodes machines
C. You don't need to restart any daemon, as they will pick up changes automatically
D. You must restart the NameNode daemon to apply the changes to the cluster
E. You must restart all six DatNode daemon to apply the changes to the cluster

35.You have installed a cluster HDFS and MapReduce version 2 (MRv2) on YARN. You have no dfs.hosts entry(ies) in your hdfs-site.xml configuration file. You configure a new worker node by setting fs.default.name in its configuration files to point to the NameNode on your cluster, and you start the DataNode daemon on that worker node. What do you have to do on the cluster to allow the worker node to join, and start sorting HDFS blocks?
A. Without creating a dfs.hosts file or making any entries, run the commands hadoop.dfsadminrefreshModes on the NameNode
B. Restart the NameNode
C. Creating a dfs.hosts file on the NameNode, add the worker Node's name to it, then issue the command hadoop dfsadmin –refresh Nodes = on the Namenode
D. Nothing; the worker node will automatically join the cluster when NameNode daemon is started

36.You use the hadoop fs –put command to add a file "sales.txt" to HDFS. This file is small enough that it fits into a single block, which is replicated to three nodes in your cluster (with a replication factor of 3). One of the nodes holding this file (a single block) fails. How will the cluster handle the replication of file in this situation?
A. The file will remain under-replicated until the administrator brings that node back online
B. The cluster will re-replicate the file the next time the system administrator reboots the NameNode daemon (as long as the file's replication factor doesn't fall below)
C. This will be immediately re-replicated and all other HDFS operations on the cluster will halt until the cluster's replication values are resorted
D. The file will be re-replicated automatically after the NameNode determines it is under-replicated based on the block reports it receives from the NameNodes

37.Given:

```
[user1@host1 ~] yarn application –list

Total Applications: 3
```

Application ID	Application-Name	Application-Type	User	Queue	State	Final-State	Progress	Tracking
Application_1374638600275_0109	Sleep Job	MAPREDUCE	user1	KILLED	KILLED	KILLED	100%	host1:54059
Application_1374638600275_0121	Sleep Job	MAPREDUCE	user1	FINISHED	SUCCEEDED	SUCCEEDED	100%	host1:19888/Jobhistory/Job_1374638600275_0121
Application_1374638600275_0020	Sleep Job	MAPREDUCE	user1	FINISHED	SUCCEEDED	SUCCEEDED	100%	host1:19888/Jobhistory/Job_1374638600275_0020

You want to clean up this list by removing jobs where the State is KILLED. What command you enter?

A. Yarn application –refreshJobHistory
B. Yarn application –kill application_1374638600275_0109
C. Yarn rmadmin –refreshQueue
D. Yarn rmadmin –kill application_1374638600275_0109

38. Assume you have a file named foo.txt in your local directory. You issue the following three commands:
Hadoop fs –mkdir input
Hadoop fs –put foo.txt input/foo.txt
Hadoop fs –put foo.txt input
What happens when you issue the third command?
A. The write succeeds, overwriting foo.txt in HDFS with no warning
B. The file is uploaded and stored as a plain file named input
C. You get a warning that foo.txt is being overwritten
D. You get an error message telling you that foo.txt already exists, and asking you if you would like to overwrite it.
E. You get a error message telling you that foo.txt already exists. The file is not written to HDFS
F. You get an error message telling you that input is not a directory
G. The write silently fails

39. You are configuring a server running HDFS, MapReduce version 2 (MRv2) on YARN running Linux. How must you format underlying file system of each DataNode?
A. They must be formatted as HDFS
B. They must be formatted as either ext3 or ext4
C. They may be formatted in any Linux file system
D. They must not be formatted - - HDFS will format the file system automatically

40. You are migrating a cluster from MApReduce version 1 (MRv1) to MapReduce version 2 (MRv2) on YARN. You want to maintain your MRv1 TaskTracker slot capacities when you migrate. What should you do/
A. Configure yarn.applicationmaster.resource.memory-mb and yarn.applicationmaster.resource.cpu-vcores so that ApplicationMaster container allocations match the capacity you require.
B. You don't need to configure or balance these properties in YARN as YARN dynamically balances resource management capabilities on your cluster
C. Configure mapred.tasktracker.map.tasks.maximum and mapred.tasktracker.reduce.tasks.maximum ub yarn-site.xml to match your cluster's capacity set by the yarn-scheduler.minimum-allocation
D. Configure yarn.nodemanager.resource.memory-mb and yarn.nodemanager.resource.cpuvcores to match the capacity you require under YARN for each NodeManager

41. On a cluster running MapReduce v2 (MRv2) on YARN, a MapReduce job is given a directory of 10 plain text files as its input directory. Each file is made up of 3 HDFS blocks. How many Mappers will run?
A. We cannot say; the number of Mappers is determined by the ResourceManager
B. We cannot say; the number of Mappers is determined by the developer
C. 30
D. 3
E. 10
F. We cannot say; the number of mappers is determined by the ApplicationMaster

42. You're upgrading a Hadoop cluster from HDFS and MapReduce version 1 (MRv1) to one running HDFS and MapReduce version 2 (MRv2) on YARN. You want to set and enforce version 1 (MRv1) to one running HDFS and MapReduce version 2 (MRv2) on YARN. You want to set and enforce a block size of 128MB for all new files written to the cluster after upgrade. What should you do?
A. You cannot enforce this, since client code can always override this value
B. Set dfs.block.size to 128M on all the worker nodes, on all client machines, and on the NameNode, and set the parameter to final
C. Set dfs.block.size to 128 M on all the worker nodes and client machines, and set the parameter to final. You do not need to set this value on the NameNode
D. Set dfs.block.size to 134217728 on all the worker nodes, on all client machines, and on the NameNode, and set the parameter to final
E. Set dfs.block.size to 134217728 on all the worker nodes and client machines, and set the parameter to final. You do not need to set this value on the NameNode

43. Your cluster has the following characteristics:
-A rack aware topology is configured and on -Replication is set to 3 -Cluster block size is set to 64MB
Which describes the file read process when a client application connects into the cluster and requests a 50MB file?
A. The client queries the NameNode for the locations of the block, and reads all three copies. The first copy to complete transfer to the client is the one the client reads as part of hadoop's speculative execution framework.
B. The client queries the NameNode for the locations of the block, and reads from the first location in the list it receives.
C. The client queries the NameNode for the locations of the block, and reads from a random location in the list it receives to eliminate network I/O loads by balancing which nodes it retrieves data from any given time.
D. The client queries the NameNode which retrieves the block from the nearest DataNode to the client then passes that block back to the client.

44. Your Hadoop cluster is configuring with HDFS and MapReduce version 2 (MRv2) on YARN. Can you configure a worker node to run a NodeManager daemon but not a DataNode daemon and still have a

functional cluster?
A. Yes. The daemon will receive data from the NameNode to run Map tasks
B. Yes. The daemon will get data from another (non-local) DataNode to run Map tasks
C. Yes. The daemon will receive Map tasks only
D. Yes. The daemon will receive Reducer tasks only

45. You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?
A. Add another master node to increase the number of nodes running the JournalNode which increases the number of machines available to HA to create a quorum
B. Set an HDFS replication factor that provides data redundancy, protecting against node failure
C. Run a Secondary NameNode on a different master from the NameNode in order to provide automatic recovery from a NameNode failure.
D. Run the ResourceManager on a different master from the NameNode in order to load-share HDFS metadata processing
E. Configure the cluster's disk drives with an appropriate fault tolerant RAID level

46. You are running Hadoop cluster with all monitoring facilities properly configured.
Which scenario will go undeselected?
A. HDFS is almost full
B. The NameNode goes down
C. A DataNode is disconnected from the cluster
D. Map or reduce tasks that are stuck in an infinite loop
E. MapReduce jobs are causing excessive memory swaps

47. You decide to create a cluster which runs HDFS in High Availability mode with automatic failover, using Quorum Storage. What is the purpose of ZooKeeper in such a configuration?
A. It only keeps track of which NameNode is Active at any given time
B. It monitors an NFS mount point and reports if the mount point disappears
C. It both keeps track of which NameNode is Active at any given time, and manages the Edits file. Which is a log of changes to the HDFS filesystem
D. If only manages the Edits file, which is log of changes to the HDFS filesystem
E. Clients connect to ZooKeeper to determine which NameNode is Active

48. Choose three reasons why should you run the HDFS balancer periodically?
A. To ensure that there is capacity in HDFS for additional data
B. To ensure that all blocks in the cluster are 128MB in size
C. To help HDFS deliver consistent performance under heavy loads

D. To ensure that there is consistent disk utilization across the DataNodes
E. To improve data locality MapReduce

49. Your cluster implements HDFS High Availability (HA). Your two NameNodes are named nn01 and nn02. What occurs when you execute the command: hdfs haadmin –failover nn01 nn02?
A. nn02 is fenced, and nn01 becomes the active NameNode
B. nn01 is fenced, and nn02 becomes the active NameNode
C. nn01 becomes the standby NameNode and nn02 becomes the active NameNode
D. nn02 becomes the standby NameNode and nn01 becomes the active NameNode

50. You have a Hadoop cluster HDFS, and a gateway machine external to the cluster from which clients submit jobs. What do you need to do in order to run Impala on the cluster and submit jobs from the command line of the gateway machine?
A. Install the impalad daemon statestored daemon, and daemon on each machine in the cluster, and the impala shell on your gateway machine
B. Install the impalad daemon, the statestored daemon, the catalogd daemon, and the impala shell on your gateway machine
C. Install the impalad daemon and the impala shell on your gateway machine, and the statestored daemon and catalogd daemon on one of the nodes in the cluster
D. Install the impalad daemon on each machine in the cluster, the statestored daemon and catalogd daemon on one machine in the cluster, and the impala shell on your gateway machine
E. Install the impalad daemon, statestored daemon, and catalogd daemon on each machine in the cluster and on the gateway node

51. You have just run a MapReduce job to filter user messages to only those of a selected geographical region. The output for this job is in a directory named westUsers, located just below your home directory in HDFS. Which command gathers these into a single file on your local file system?
A. Hadoop fs –getmerge –R westUsers.txt
B. Hadoop fs –getemerge westUsers westUsers.txt
C. Hadoop fs –cp westUsers/* westUsers.txt

D. Hadoop fs –get westUsers westUsers.txt

52. In CDH4 and later, which file contains a serialized form of all the directory and files inodes in the filesystem, giving the NameNode a persistent checkpoint of the filesystem metadata?
A. fstime
B. VERSION
C. Fsimage_N (where N reflects transactions up to transaction ID N)
D. Edits_N-M (where N-M transactions between transaction ID N and transaction ID N)

53. You are running a Hadoop cluster with a NameNode on host mynamenode. What are two ways to determine available HDFS space in your cluster?
A. Run hdfs fs –du / and locate the DFS Remaining value
B. Run hdfs dfsadmin –report and locate the DFS Remaining value
C. Run hdfs dfs / and subtract NDFS Used from configured Capacity
D. Connect to http://mynamenode:50070/dfshealth.jsp and locate the DFS remaining value

54. You have recently converted your Hadoop cluster from a MapReduce 1 (MRv1) architecture to MapReduce 2 (MRv2) on YARN architecture. Your developers are accustomed to specifying map and reduce tasks (resource allocation) tasks when they run jobs: A developer wants to know how specify to reduce tasks when a specific job runs. Which method should you tell that developers to implement?
A. MapReduce version 2 (MRv2) on YARN abstracts resource allocation away from the idea of “tasks” into memory and virtual cores, thus eliminating the need for a developer to specify the number of reduce tasks, and indeed preventing the developer from specifying the number of reduce tasks.
B. In YARN, resource allocations is a function of megabytes of memory in multiples of 1024mb. Thus, they should specify the amount of memory resource they need by executing –D mapreducereduces.memory-mb-2048
C. In YARN, the ApplicationMaster is responsible for requesting the resource required for a specific launch. Thus, executing –D yarn.applicationmaster.reduce.tasks=2 will specify that the ApplicationMaster launch two task contains on the worker nodes.
D. Developers specify reduce tasks in the exact same way for both MapReduce version 1 (MRv1) and MapReduce version 2 (MRv2) on YARN. Thus, executing –D mapreduce.job.reduces-2 will specify reduce tasks.
E. In YARN, resource allocation is function of virtual cores specified by the ApplicationManager making requests to the NodeManager where a reduce task is handeled by a single container (and thus a single virtual core). Thus, the developer needs to specify the number of virtual cores to the NodeManager by executing –pyarn.nodemanager.cpu-vcores=2

55. Your Hadoop cluster contains nodes in three racks. You have not configured the dfs.hosts property in the NameNode’s configuration file. What results?

A. The NameNode will update the dfs.hosts property to include machines running the DataNode daemon on the next NameNode reboot or with the command dfsadmin –refreshNodes
B. No new nodes can be added to the cluster until you specify them in the dfs.hosts file
C. Any machine running the DataNode daemon can immediately join the cluster
D. Presented with a blank dfs.hosts property, the NameNode will permit DataNodes specified in mapred.hosts to join the cluster

56. You are running a Hadoop cluster with MapReduce version 2 (MRv2) on YARN. You consistently see that MapReduce map tasks on your cluster are running slowly because of excessive garbage collection of JVM, how do you increase JVM heap size property to 3GB to optimize performance?
A. yarn.application.child.java.opts=-Xsx3072m
B. yarn.application.child.java.opts=-Xmx3072m
C. mapreduce.map.java.opts=-Xms3072m
D. mapreduce.map.java.opts=-Xmx3072m

57. You have a cluster running with a FIFO scheduler enabled. You submit a large job A to the cluster, which you expect to run for one hour. Then, you submit job B to the cluster, which you expect to run a couple of minutes only.
You submit both jobs with the same priority.
Which two best describes how FIFO Scheduler arbitrates the cluster resources for job and its tasks?
A. Because there is a more than a single job on the cluster, the FIFO Scheduler will enforce a limit on the percentage of resources allocated to a particular job at any given time
B. Tasks are scheduled on the order of their job submission
C. The order of execution of job may vary
D. Given job A and submitted in that order, all tasks from job A are guaranteed to finish before all tasks from job B
E. The FIFO Scheduler will give, on average, and equal share of the cluster resources over the job lifecycle
F. The FIFO Scheduler will pass an exception back to the client when Job B is submitted, since all slots on the cluster are use

58.A user comes to you, complaining that when she attempts to submit a Hadoop job, it fails. There is a Directory in HDFS named /data/input. The Jar is named j.jar, and the driver class is named DriverClass. She runs the command:
Hadoop jar j.jar DriverClass /data/input/data/output
The error message returned includes the line:
PriviligedActionException as:training (auth:SIMPLE)
cause:org.apache.hadoop.mapreduce.lib.input.invalidInputException:
Input path does not exist: file:/data/input
What is the cause of the error?

A. The user is not authorized to run the job on the cluster
B. The output directory already exists
C. The name of the driver has been spelled incorrectly on the command line
D. The directory name is misspelled in HDFS
E. The Hadoop configuration files on the client do not point to the cluster

59.Your company stores user profile records in an OLTP databases. You want to join these records with web server logs you have already ingested into the Hadoop file system. What is the best way to obtain and ingest these user records?
A. Ingest with Hadoop streaming
B. Ingest using Hive's IQAD DATA command
C. Ingest with sqoop import
D. Ingest with Pig's LOAD command
E. Ingest using the HDFS put command

60.Which two are features of Hadoop's rack topology?
A. Configuration of rack awareness is accomplished using a configuration file. You cannot use a rack topology script.
B. Hadoop gives preference to intra-rack data transfer in order to conserve bandwidth
C. Rack location is considered in the HDFS block placement policy
D. HDFS is rack aware but MapReduce daemon are not
E. Even for small clusters on a single rack, configuring rack awareness will improve performance